تمرین سری 1 زهرا عبادی

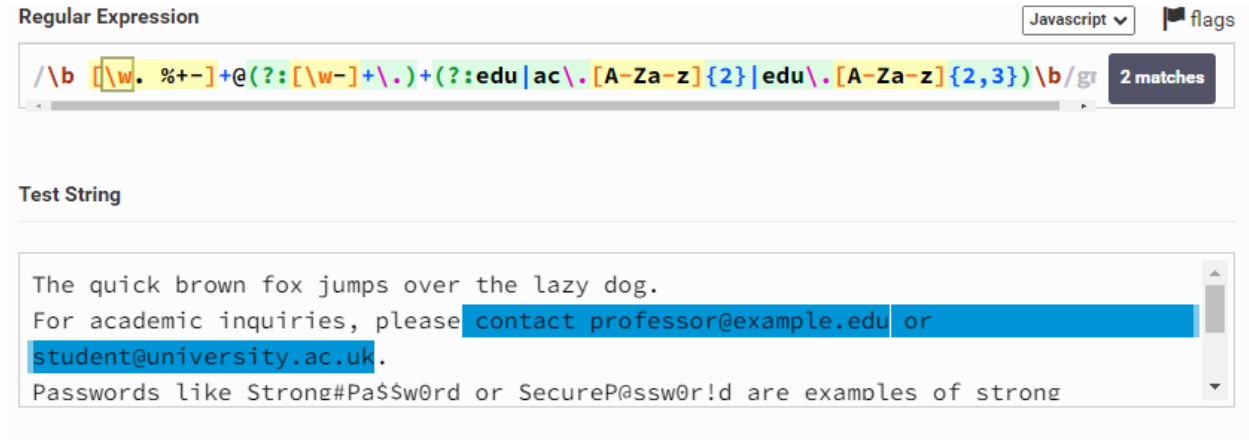
.1

الف)

^[A-Z]

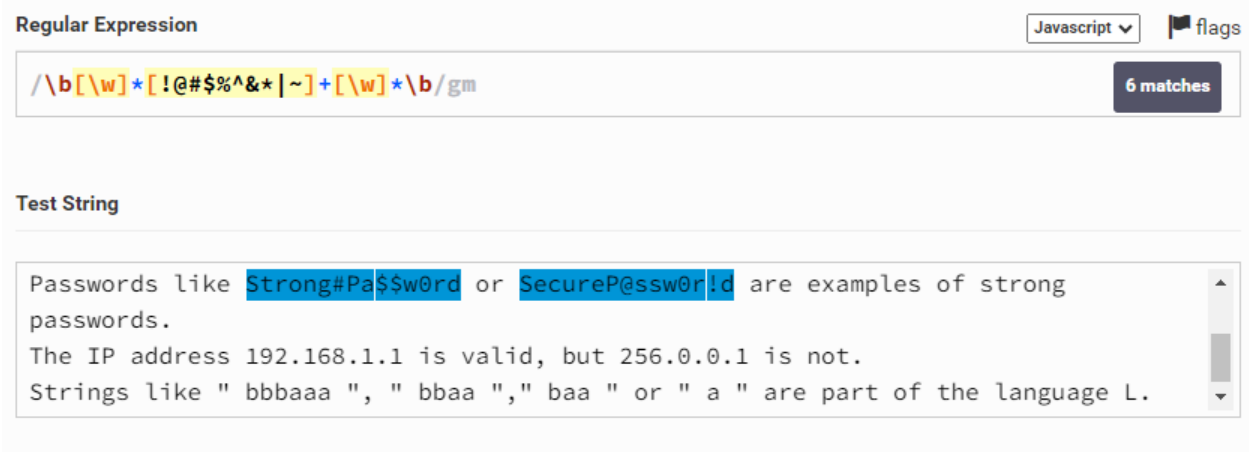
ب)

\b[\w. %+-]+@(?:[\w-]+\.)+(?:edu|ac\.[A-Za-z]{2}|edu\.[A-Za-z]{2,3})\b



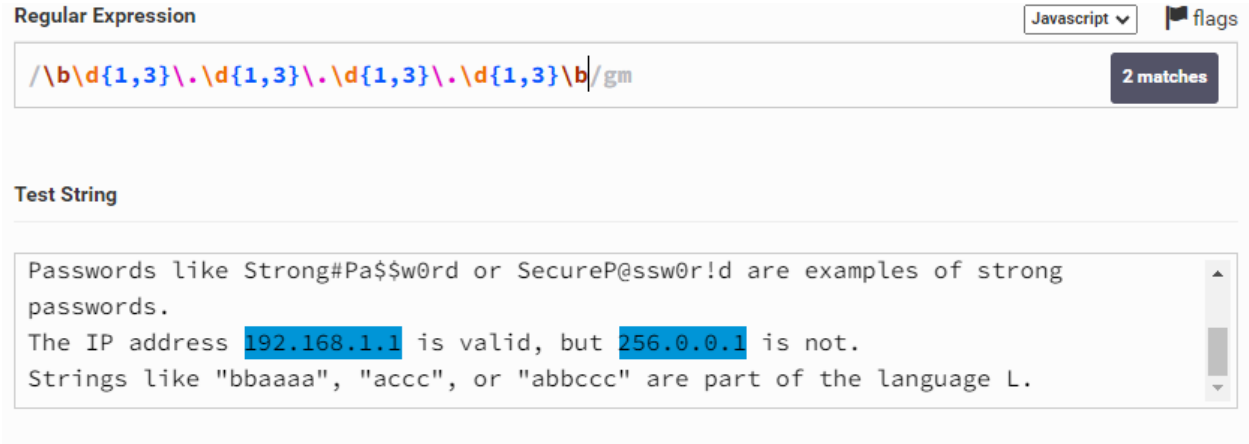
ج)

\b[\w]*[!@#$%^&*|~]+[\w]*\b



د)

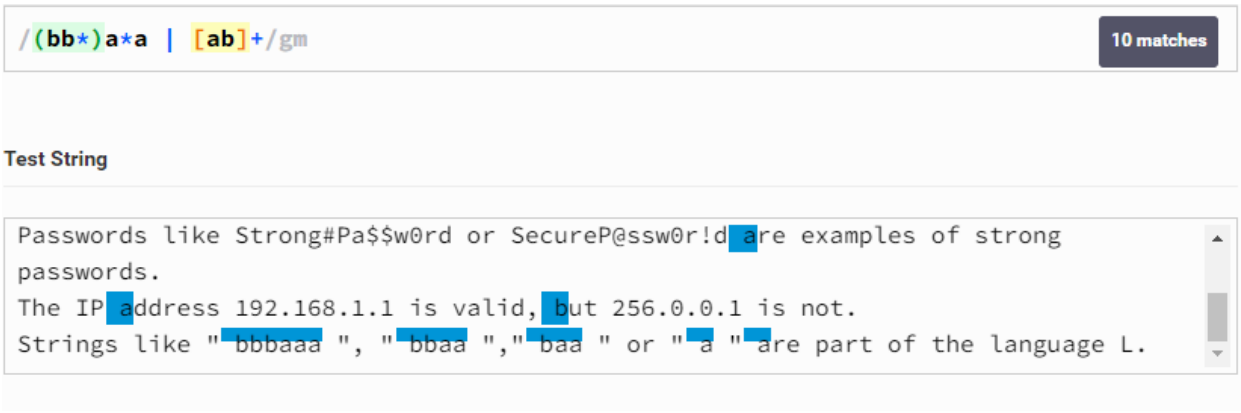
\b\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}\b



ه)

bbbaaa
bbaa
baa
a
b

(bb*)a*a | [ab]+



The quick brown fox jumps over the lazy dog.
For academic inquiries, please contact professor@example.edu or student@university.ac.uk.
Passwords like Strong#Pa$$w0rd or SecureP@ssw0r!d are examples of strong passwords.
The IP address 192.168.1.1 is valid, but 256.0.0.1 is not.
Strings like " bbbaaa ", " bbaa "," baa " or " a " are part of the language L.

2. پیکره بی جن خان، در آزمایشگاه زبان شناسی دانشگاه تهران نگهداری می شود. این پیکره، از برخی اخبار روزنامه ها و متون معمولی جمع آوری شده است. یکی از ویژگی های این پیکره این است که هر سند در این مجموعه دارای یک عنوان می باشد. به عنوان مثال، اسناد تحت عناوین(سیاسی، فرهنگی، اقتصادی)دسته بندی شده اند.در این پیکره **4300** عنوان مختلف وجود دارد. این عنوان ها یک محیط آزمایشی مورد دلخواه برای خوشه بندی و مقوله بندی و غیره را تولید می کند. این پیکره شامل **2598215** واژه و **550** برچسب می باشد که به طور دستی برچسب زده شده است. در عملیات برچسب زنی از عناوین متون صرف نظر شده است. زیرا هدف، بدست آوردن یک نرم افزار برچسب زننده خودکار است

هر برچسب در این مجموعه از یک ساختار سلسله مراتبی پیروی می کند. بخشهایی از نام برچسب که در ابتدای نام آن قرار دارند، بیان کننده توصیف کلی تری از آن برچسب می باشند. در ابتدای برچسب مقوله های اصلی مشخص می شوند، بخشهایی که در انتهای نام برچسب قراردارند، توصیف جزئی تر در مورد آن برچسب هستند. یعنی سایر ویژگی های مقوله های اصلی قرار می گیرند. مثال برچسب**N_PL_LOC** سه سطح در ساختار سلسله مراتبی می باشد. سطح اول

Nمشخص کننده اسم می باشد. سطح دوم**PL** مشخص کننده نوع جمع می باشد و سطح سوم **LOC** مشخص کننده مکان می باشد.

پیکره درختی وابستگی فارسی اوپسالا (**UPDT**) مجموعه‌ای است از جملات فارسی که در آن روابط نحوی کلمات بر مبنای دستور وابستگی مشخص شده است. این پیکره که در دانشگاه اوپسالای سوئد تهیه شده است، حاوی **6000** جمله برگرفته از پیکره فارسی اوپسالا (**UPC** – نسخه‌ای تغییریافته از پیکره بی‌جن‌خان) می‌باشد و بر اساس قالب **conll** و طرح برچسب‌زنی Stanford Typed Dependencies تهیه شده است. ناشر این پیکره دپارتمان زبان‌شناسی و فیلولوژی، دانشگاه اوپسالا، سوئد است.

3. WordNetو **FarsNet** پایگاه‌های واژگانی هستند که کلمات را در روابط معنایی سازماندهی می‌کنند و اطلاعات ارزشمندی درباره معانی کلمات، مترادف‌ها، متضادها و سایر ویژگی‌های واژگانی ارائه می‌دهند. در اینجا یک مرور کلی از هر یک و برنامه های آنها آورده شده است :

1. WordNet :

WordNet یک پایگاه داده واژگانی از اسامی، افعال، صفت ها و قیدهای انگلیسی است که به صورت **synsets** (مجموعه ای از مترادف ها) سازماندهی شده اند. هر synset یک مفهوم متمایز را نشان می دهد و از طریق روابط معنایی مختلف مانند hypernym ها (اصطلاحات وسیع تر)، هیپونیم ها (اصطلاحات محدودتر)، مرونیم ها (روابط جزئی-کل) و هولونیم ها (روابط کل جزئی) به مجموعه های دیگر مرتبط می شود.

برنامه های کاربردی:

- پردازش زبان طبیعی **(NLP) WordNet** :به طور گسترده در کارهایی مانند طبقه بندی متن، تجزیه و تحلیل احساسات و بازیابی اطلاعات استفاده می شود.
- ترجمه ماشینی: با ارائه معادل های معنایی برای کلمات در زبان های مختلف به بهبود دقت سیستم های ترجمه ماشینی کمک می کند.
- بازیابی اطلاعات **WordNet** :با گرفتن روابط معنایی بین کلمات، به نمایه سازی و بازیابی اسناد کمک می کند و دقت جستجو را افزایش می دهد.

مثال:

کلمه "سگ" را در نظر بگیرید. در**WordNet** ، آن را به‌عنوان یک synset همراه با مترادف‌های آن (به عنوان مثال، "سنگ") و مفاهیم مرتبط مانند ابرنام (به عنوان مثال، "حیوان")، مترادف (به عنوان مثال، "توله سگ") و مترونیم‌ها (به عنوان مثال، "دم" نشان داده می‌شود. ")

1. FarsNet :

FarsNet یک پایگاه واژگانی مشابه WordNet است اما به طور خاص برای زبان فارسی (فارسی) طراحی شده است. حاوی اطلاعات معنایی در مورد کلمات فارسی از جمله مترادف، متضاد، ابرنام، مترادف و سایر روابط معنایی است.

برنامه های کاربردی:

- پردازش زبان فارسی **FarsNet** : برای کارهای مختلف **NLP** در زبان فارسی، از جمله تجزیه و تحلیل احساسات، تشخیص موجودیت نام‌گذاری شده و خلاصه‌سازی متن بسیار مهم است.
- کاربردهای بین زبانی: با ارائه معادل های معنایی، بازیابی اطلاعات بین زبانی و ترجمه ماشینی بین فارسی و سایر زبان ها را تسهیل می کند.
- فرهنگ نویسی و آموزش زبان: فارس نت به عنوان منبعی ارزشمند برای فرهنگ نویسان، معلمان زبان و زبان آموزان است که به گسترش و درک واژگان کمک می کند.

مثال:

کلمه فارسی «خرس» (خرس) را در نظر بگیرید که به معنای «خرس» است. در فارس نت، این کلمه به صورت ترکیبی همراه با مترادف‌ها (مثلاً «خرس قهوه‌ای» برای خرس قهوه‌ای)، متضادها (مثلاً «پری»)، ابرنام‌ها (مثلاً «جانور») و سایر مفاهیم مرتبط نشان داده می‌شود.

به طور خلاصه، **WordNet** و **FarsNet** هر دو نقش تعیین کننده ای در **NLP**، فناوری زبان و تحقیقات زبانی با ارائه اطلاعات معنایی غنی در مورد کلمات به ترتیب به زبان انگلیسی و فارسی ایفا می کنند.

4. ابزار **NLTK** (طبیعی پردازش زبان) یکی از پرکاربردترین و مفیدترین ابزارهای استفاده شده در زمینه پردازش زبان‌های طبیعی است. این ابزار دارای مجموعه‌ای از دیتاست‌ها و ابزارهاست که به تحلیل، پردازش و استخراج اطلاعات از متون مختلف کمک می‌کند. در زیر، سه مورد از دیتاست‌ها و سه مورد از ابزارهای مهم **NLTK** را معرفی و کاربردهای آن‌ها را توضیح می‌دهم :

دیتاست‌ها:

1. **دیتاست WordNet :**

- WordNetیک دیتاست معروف در زمینه‌ی هم‌معنی‌ها(synonyms) ، هم‌زمان‌ها (antonyms) و روابط معنایی بین کلمات است.
- کاربردهای آن شامل پردازش معنایی کلمات، مترادف‌یابی، ساختاردهی و مدل‌سازی زبانی است.

2. دیتاست **Gutenberg Corpus** :

- این دیتاست شامل مجموعه‌ای از متون کلاسیک و عمومی است که از پروژه Gutenberg برداشته شده است.
- کاربردهای آن شامل آموزش مدل‌های زبانی، تحلیل متون تاریخی و ادبیاتی، و ایجاد مجموعه‌های آموزشی و آزمونی است.

3. دیتاست **Movie Reviews :**

- این دیتاست شامل بررسی‌ها و نقدهای کاربران در مورد فیلم‌ها است.
- کاربردهای آن شامل ایجاد سیستم‌های تشخیص متن و احساسات(Sentiment Analysis) ، تحلیل دیدگاه‌ها و ارزیابی کیفیت فیلم‌هاست.

ابزارها:

1. **Tokenizer** :

- این ابزار برای تجزیه متن به تکه‌های کوچکتر مانند کلمات یا جملات استفاده می‌شود.
- کاربردهای آن شامل تجزیه و تحلیل متن، پردازش زبانی، و پیش‌پردازش متون برای وظایف دیگر مانند تحلیل عاطفه است.

2. **POS Tagger (Part-of-Speech Tagger)** :

- این ابزار برای تشخیص نقش هر کلمه در یک جمله (مثلا فعل، اسم، صفت و ...) استفاده می‌شود.
- کاربردهای آن شامل استخراج اطلاعات ساختار جملات، تحلیل گرامری، و ترجمه ماشینی است.

3. **Named Entity Recognizer (NER)** :

- این ابزار برای شناسایی و استخراج اسامی نهادها، مکان‌ها، افراد و سایر موجودیت‌های مهم از متن استفاده می‌شود.
- کاربردهای آن شامل استخراج اطلاعات از متون، تحلیل اخبار، و ایجاد پایگاه داده‌های نام‌گذاری شده است.